**STEPS IN COMPLETING CHARGE GLI PROJECT META-ANALYSIS (PHASE II)**

*Authors: Thomas Winkler, Karen Schwander, Amy Bentley, Michael Brown, Raymond Noordam, Songmi Lee, Pavithra Nagarajan, Heming Wang, Hugues Aschard, Alisa Manning, Jim Gauderman, DC Rao*

*Contact: Thomas Winkler (*[*thomas.winkler@ukr.de*](mailto:thomas.winkler@ukr.de)*), DC Rao (*[*rao@wustl.edu*](mailto:rao@wustl.edu)*)*

*Note: The below steps are guidelines, and project teams should make whatever adjustments are the most appropriate based on their data and own judgment.*

*Latest version available from:* [*https://wustl.app.box.com/s/5m1lbx8wpdife1kg32bxg81voh4kapvg/folder/131454415489*](https://wustl.app.box.com/s/5m1lbx8wpdife1kg32bxg81voh4kapvg/folder/131454415489) *(Subfolder: “/Analysis Plans/0.Meta-Analysis”)*

# Material

* *For QC: <To Be Included>*
* *For Meta-analyses: Download METAL from* [*https://genome.sph.umich.edu/wiki/Meta\_Analysis\_of\_SNPxEnvironment\_Interaction*](https://genome.sph.umich.edu/wiki/Meta_Analysis_of_SNPxEnvironment_Interaction)
* *For Approaches: Download R package EasyStrata2, EasyStrata2 scripts and reference files (for PCA and clumping) from* [*https://www.genepi-regensburg.de/charge-gli*](https://www.genepi-regensburg.de/charge-gli)
  + *An overview of the EasyStrata2 scripts that implement our post-meta analytical framework (i.e., the bidirectional and combined-meta approaches) can be found in* ***Supplementary Table 1***

# Protocol

## Expected input

*The protocol expects study-specific genome-wide interaction study (GWIS) and genome-wide association study (GWAS) results as input that are based on the following regression models:*

* *“Model 1”:*
* *“Model 2”:*

## Study-Ancestry QC

1. Run EasyQC2 on each Study-Ancestry Group
   1. Double-check that no analyses have been submitted for strata with n<100 individuals. Although this is specified in APs, some have been uploading anyway. Exclude these from all further steps.
   2. Traits of the same type (quantitative and/or categorical) can be run together
      1. Quantitative: edit out N\_exp, MAC\_0, MAC\_1, EAF\_0, EAF\_1 references.
   3. Different ancestries must be run separately
      1. Edit MERGE statement for allele frequency reference files
2. Evaluate QC output files and troubleshoot
   1. Allele Frequency Plots: expectation is that most points fall along the diagonal in comparison with an ancestry-appropriate reference, with width of scatter variable based on how well ancestry of participants matches available reference.
   2. QQ plots
   3. Rep files: summarize cleaning steps (useful to compare across studies)

## Ancestry QC

1. Meta QC – EasyQC run on all CLEANED files within an ancestry/trait/exposure group.
   1. Comparison of plots using different filtering strategies
      1. At minimum, use df ≥ 20 [minimum(MAC0,MAC1)\*imputation score ≥ 20] (higher at project team’s discretion)
      2. Imputation score ≥ 0.5
   2. SEN plots useful to make sure no errors with trait transformation, E centering, etc.
      1. Expectation is that points (representing cohorts) lie more or less along diagonal. Any outlier should be investigated.
   3. QQ plots [currently not included in metaQC]
   4. Check for E-centering
      1. Outlying studies nearing y=0 may have used centered results
   5. Rep files provide details on the cleaned files for comparisons across included studies and should be evaluated for outliers. For instance, outliers in beta distribution might indicate problems with trait transformation.
   6. NOTE: MetaQC does not produce filtered results files, but gives plots related to each filtering strategy to enable decision-making on filters by the meta-analyst. Filtering decisions should be implemented in meta-analysis.

## Determine subset composition

1. Once all study data has been cleaned, divide the studies into two subsets that are roughly equivalent in sample size, which will now make up Stage 1 (**S1**) and Stage 2 (**S2**). NOTE: the goal is to achieve relative balance between S1 and S2 sample sizes within population/trait/exposure analysis units. It is not necessary that the divisions be the same across analyses (e.g. ARIC AFR could be in S1 for the AFR meta-analysis while ARIC EUR are in S1 for the EUR meta-analysis).

## Population-specific meta-analyses

1. In METAL, pool population-specific study results (**Supplementary Figure 1**)
   1. Example input: “Study1\_EUR”, “Study2\_EUR”, … “StudyM\_EUR”
   2. Run separately by population.
   3. Run separately for S1 and S2 and on all studies combined (S1plusS2).
   4. Include filtering strategy determined from Population QC using ADDFILTER:
      1. For quantitative E: ADDFILTER DF\_ALL>=20
      2. For binary E: ADDFILTER DF\_ALL>=20, ADDFILTER DF\_E0>=20 and ADDFILTER DF\_E1>=20
   5. Use the genomic control option (i.e., Single GC, “study-level”).
   6. Use the heterogeneity option (to derive between-study heterogeneity estimates and HetDf for filtering on the number of studies later in this process).
   7. Run separately for
      1. Model 1 (joint meta-analysis): SCHEME INTERACTION (METAL Add-on for interaction needed; this will produce the joint 2df P-Value for identification of novel loci; please discard any main and interaction effect estimates produced)
      2. Model 1 (interaction effect meta-analysis): SCHEME STDERR (this will produce the 1 df interaction effect estimates and interaction P-Values for interpretation and identification of interaction loci)
      3. Model 2 (marginal effect meta-analysis): SCHEME STDERR (this will produce the 1 df marginal P-value that is required for the 2-step approach)
   8. Expected results: Single-GC corrected population-specific metal results for S1, S2 and S1plusS2 (e.g., “S1\_EUR\_1GC”, “S1\_AFR\_1GC”,…, “S2\_EUR\_1GC”, “S2\_AFR\_1GC”, …, “S1plusS2\_EUR\_1GC”, “S1plusS2\_AFR\_1GC”)

## Cross-population meta-analyses

1. In METAL, pool population-specific metal results (from 5.; **Supplementary Figure 2**)
   1. Example input: “S1\_EUR\_1GC”, “S1\_AFR\_1GC” and “S1\_EAS\_1GC”
   2. Run separately for S1, S2 and on all studies combined (S1plusS2).
   3. Use the genomic control option (i.e., Double GC, “population-level”).
   4. Use the heterogeneity option (to derive between-ancestry heterogeneity estimates).
   5. Run separately for Model 1 (joint meta-analysis, SCHEME INTERACTION), Model 1 (interaction effect meta-analysis, SCHEME STDERR) and Model 2 (marginal effect meta-analysis, SCHEME STDERR)
   6. Expected results: Double-GC corrected cross-population metal results for S1, S2 and S1plusS2 (“S1\_ALL\_2GC”, “S2\_ALL\_2GC” and “S1plusS2\_ALL\_2GC”)

## Bi-directional approaches on S1 and S2 results

1. Bi-directional approach on population-specific S1 and S2 metal results (from 5.)
   1. Example input: “S1\_EUR\_1GC” and “S2\_EUR\_1GC”
   2. Run separately by population.
   3. Do NOT run bi-directional if total sample size is <20,000 in S1 or S2. Lower thresholds are up to the project teams.
   4. EasyStrata2 scripts “bidirectional\_popspecific\_joint2df.ecf” and “bidirectional\_popspecific\_ int1df.ecf” implement the following steps e-i.
   5. Preprocessing
      1. Remove missing values (e.g., variants that are only included in one stage)
      2. Remove low frequency variants MAF<0.001
      3. Obtain chromosome and position from the MarkerName (e.g., 1:3257725:A\_C)
      4. Remove MHC region variants +/-1Mb (chr6:27500000-34500000, <https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC>)
      5. Calculate interaction P-Values:

*Pint\_stage1 = 2\*pnorm(abs(IntEffect\_stage1/IntStdErr\_stage1),lower.tail=F)*

*Pint\_stage2 = 2\*pnorm(abs(IntEffect\_stage2/IntStdErr\_stage2),lower.tail=F)*

* + 1. Calculate GC lambdas and apply GC correction to marginal, interaction and joint2df P-Values (i.e., Double GC, “population-level”)
    2. Calculate 1-sided interaction P-Values depending on the direction of the interaction effect in the other stage:

*Pint1sided\_stage1.GC = pnorm(sign(IntEffect\_stage2)\*IntEffect\_stage1/IntStdErr\_stage1.GC,lower.tail=F)*

*Pint1sided\_stage2.GC = pnorm(sign(IntEffect\_stage1)\*IntEffect\_stage2/IntStdErr\_stage2.GC,lower.tail=F)*

* 1. Test for 1df GxE: S1 for Discovery, S2 for Replication
     1. In S1: Test for 1df GxE (mGxE variants)
        1. Test for 1df GxE (Model 1: PGxE ≤ 5x10-8), and
        2. 2-step GxE:
           1. Step 1: Subset on 1df G (Model 2: PG ≤ 10-5, mG variants)
           2. Step 2: Test for 1df GxE (Model 1: PGxE ≤ 0.05/[effective number of tests by PCA on mG variants])
        3. Derive mGxE variantsas the union of significant variants obtained from “1.” and “2.”.
     2. In S2: Test mGxE variants for 1df GxE (Model 1: One-sided PGxE ≤ 0.05/2/[effective number of tests by PCA on mGxE variants])
     3. Clump significant variants (d>500kb, r2<0.1) for independent locus lead variants
  2. Repeat f. but with S2 for Discovery, S1 for Replication
  3. Test for joint 2df: S1 for Discovery, S2 for Replication
     1. In S1: Test for joint 2df (Model 1: P2df ≤ 5x10-8) (m2df variants)
     2. In S2: Test m2df variants for joint 2df (Model 1: P2df ≤ 0.05/2/[effective number of tests by PCA on m2df variants])
     3. Clump significant variants (d>500kb, r2<0.1) for independent locus lead variants.
  4. Repeat h. but with S2 for Discovery, S1 for Replication

1. Bi-directional approach on cross-population S1 and S2 metal results (from 6.)
   1. Example input: “S1\_ALL\_2GC” and “S2\_ALL\_2GC”
   2. Do NOT run bi-directional if total sample size is <20,000 in S1 or S2. Lower thresholds are up to the project teams.
   3. Repeat steps e.-i. from 7. but omit the GC correction (!!!)
   4. EasyStrata2 scripts “bidirectional\_crosspop\_joint2df.ecf” and “bidirectional\_ crosspop\_int1df.ecf” implement the approach.

## Combined-meta approaches on S1plusS2 results

1. Combined-meta approach on population-specific S1plusS2 metal results (from 5.)
   1. Example input: “S1plusS2\_EUR\_1GC”
   2. Run separately by population.
   3. Do NOT run combined-meta approaches if total sample size is <20,000 or the number of studies was <3. Lower thresholds are up to the project teams.
   4. EasyStrata2 scripts “combmeta\_popspecific\_joint2df.ecf” and “combmeta\_popspecific\_int1df.ecf” implement the following steps e- g.
   5. Preprocessing
      1. Remove missing values
      2. Remove low frequency variants MAF<0.001
      3. Obtain chromosome and position from the MarkerName (e.g., 1:3257725:A\_C)
      4. Remove MHC region variants (chr6:27500000-34500000, <https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC>)
      5. Calculate GC lambdas and apply GC correction to marginal, interaction and joint2df P-Values (Double ancestry-level GC)
      6. Calculate interaction P-Values:

*Pint = 2\*pnorm(abs(IntEffect/IntStdErr),lower.tail=F)*

* 1. Test for 1df GxE
     1. Calculate 1df GxE FDR based on all variants genome-wide (FDRGxE)
     2. Test for 1df GxE (Model 1: ~~P~~~~GxE~~ ~~≤ 5x10~~~~-9~~ ~~and FDR~~~~GxE~~ ~~<0.05~~ PGxE ≤ 5x10-8)
     3. 2-step GxE:
        1. Step 1: Subset on 1df G (Model 2: PG ≤ 10-5, mG variants)
        2. Calculate 1df GxE FDR based on mG variants (FDRGxE,mG)
        3. Step 2: Test for 1df GxE (Model 1: PGxE ≤ 0.05/[effective number of tests by PCA on mG variants] and FDRGxE,mG<0.05)
     4. Clump significant variants (d>500kb, r2<0.1) for independent locus lead variants
  2. Test for joint 2df
     1. Calculate 2df joint FDR based on all variants genome-wide (FDR2df)
     2. Test for 1df GxE (Model 1: ~~P~~~~2df~~ ~~≤ 5x10~~~~-9~~ ~~and FDR~~~~2df~~ ~~<0.05~~ P2df ≤ 5x10-8)
     3. Clump significant variants (d>500kb, r2<0.1) for independent locus lead variants

1. Combined-meta approach on cross-population S1plusS2 metal results (from 6.)
   1. Example input: “S1plus2\_ALL\_2GC”
   2. Do NOT run combined-meta if total sample size was <20,000 and number of studies is <3. Lower thresholds are subject to the project teams.
   3. Repeat steps e.-g. from 9. but omit the GC correction (!!!)
   4. EasyStrata2 scripts “combmeta\_crosspop\_joint2df.ecf” and “combmeta\_ crosspop\_int1df.ecf” implement the approach.

## Post-Approaches Processing

1. Filtering on # of studies included (using hetdf column from the METAL output file)
   1. Further discussion needed. Phase I: 3 studies (though lower for some population groups with few studies total)
   2. Note: HetDf should be 1 larger than the # of studies filter for 1df and main effect models, 2 larger than the # of studies filter for 2df studies
      1. Can also add an N\_Cohort column to the files by counting non-? Characters in the Direction column of metal output.
2. Filtering on sample size
   1. Bi-directional analyses: N ≥ 20,000 in S1 and S2
   2. Combined analyses: number of studies ≥ 3 OR N ≥ 20,000
   3. Lower thresholds for population groups with a small number of contributing studies to be set at the project team’s discretion.
3. Separating known from novel loci
   1. EasyStrata can read in a known loci list and produce Manhattan plots with those colored differently and QQ plots with and without those loci
   2. d>500 kb or r2<0.1 to known loci previously used to determine whether variants are “known” or “novel”.
      1. EasyStrata can group based on distance and LD using INDEP function
   3. LD-based clumping

# Supplementary Material

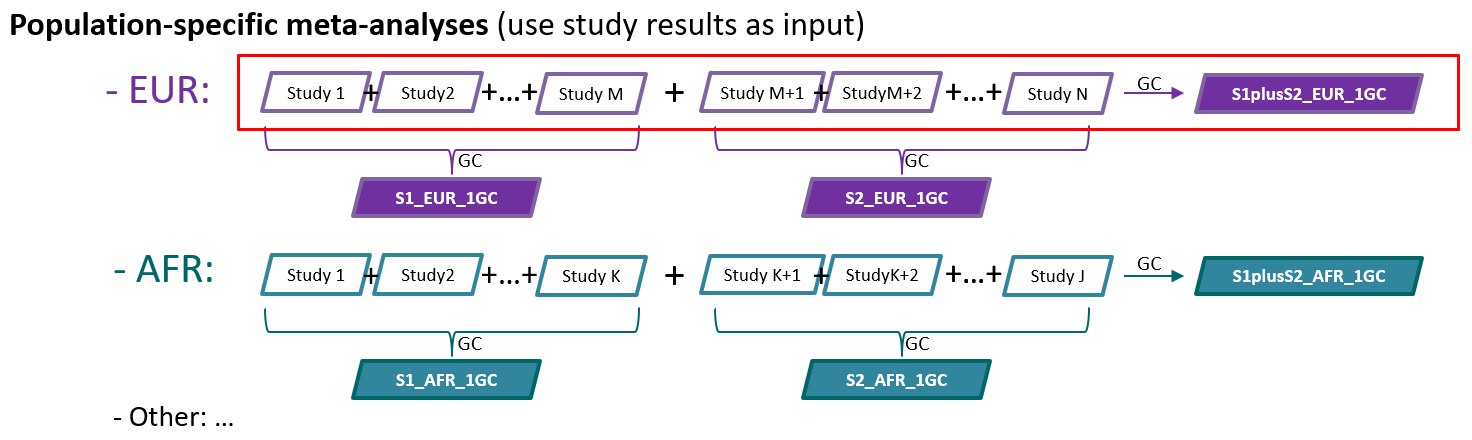
## Supplementary Table 1. Overview on EasyStrata2 scripts.

The table shows an overview on the eight EasyStrata2 scripts grouped by approach, population and test. Any population-specific metal results as input are expected to be single-GC corrected (i.e., study-level GC corrected) and any cross-population metal results as input are expected to be double-GC corrected (i.e., study- and ancestry-level GC corrected). Each script can be run directly on the metal results and will produce lists of significant independent region lead variants (d>500kb, see output file indicated by “\*.indep.regionleads.txt”) as well as independent index variants (r2<0.1 within regions, see output file indicated by “\*.indep.locusleads.txt”).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **Population** | **Test** | **Script name** | **Example input** |
| Bidirectional | Population-specific | Joint 2df | bidirectional\_popspecific\_joint2df.ecf | S1\_EUR\_1GC + S2\_EUR\_1GC |
| Int 1df | bidirectional\_popspecific\_int1df.ecf |
| Cross-population | Joint 2df | bidirectional\_crosspop\_joint2df.ecf | S1\_ALL\_2GC + S2\_ALL\_2GC |
| Int 1df | bidirectional\_crosspop\_int1df.ecf |
| Combined-meta | Population-specific | Joint 2df | combmeta\_popspecific\_joint2df.ecf | S1plus2\_EUR\_1GC |
| Int 1df | combmeta\_popspecific\_int1df.ecf |
| Cross-population | Joint 2df | combmeta\_crosspop\_joint2df.ecf | S1plus2\_ALL\_2GC |
| Int 1df | combmeta\_crosspop\_int1df.ecf |

## Supplementary Figure 1. Population-specific meta-analyses.

Any population-specific meta-analyses (S1, S2 or S1plus2) should be conducted based on (non-GC corrected) study-specific results. Genomic control correction should be applied during the metal analysis (GENOMICCONTROL ON option in metal) to derive single-GC corrected population-specific meta-analysis results. Between-study heterogeneity should be estimated (ANALYZE HETEROGENEITY option in metal).



## Supplementary Figure 2. Cross-population meta-analyses.

Any cross-population meta-analyses (S1, S2 or S1plus2) should be conducted based on single-GC corrected population-specific metal results. Genomic control correction should be applied during the metal analysis (GENOMICCONTROL ON option in metal) to derive double-GC corrected cross-population meta-analysis results. Between-ancestry heterogeneity should be estimated (ANALYZE HETEROGENEITY option in metal).

